Data augmentation in food science: synthesising spectroscopic data of vegetable oils for performance enhancement

Georgouli, K., Osorio, M. T., Martinez del Rincon, J., & Koidis, A. (2018). Data augmentation in food science: synthesising spectroscopic data of vegetable oils for performance enhancement. *Journal of Chemometrics*, *32*(6), [e3004]. https://doi.org/10.1002/cem.3004

Published in:
Journal of Chemometrics

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Data augmentation in food science: synthesising spectroscopic data of vegetable oils for performance enhancement

Konstantia Georgouli[a], Maria Teresa Osorio[a], Jesus Martinez Del Rincon[b] and Anastasios Koidis[a,*]

[a] *Queens University Belfast, Institute for Global Food Security, Belfast, Northern Ireland, UK*

[b] *Queens University Belfast, Institute of Electronics, Communications and Information Technology, Belfast, Northern Ireland, UK*

Abstract

Generating more accurate, efficient and robust classification models in chemometrics, able to address real-world problems in food analysis, is intrinsically related with the amount of available calibration samples. In this paper, we propose a data augmentation solution in order to increase the performance of a classification model by generating realistic data augmented samples. The feasibility of this solution has been evaluated on three main different experiments where Fourier transform mid infrared (FT-IR) spectroscopic data of vegetable oils were used for the identification of vegetable oil species in oil admixtures. Results demonstrate that data augmented samples improved the classification rate by around 19% in a single instrument validation and provided a significant 38% improvement in classification when testing in more than 10 different spectroscopic instruments to the calibration one.

Keywords: data augmentation; artificial samples; classification; vegetable oils; spectroscopy

*Corresponding author

Email addresses: kgeorgouli01@qub.ac.uk (Konstantia Georgouli), t.osorio@qub.ac.uk (Maria T eresa Osorio), j.martinez-del-rincon@qub.ac.uk (Jesus Martinez Del Rincon), t.koidis@qub.ac.uk (Anastasios Koidis)

1. Introduction

The application of chemometrics in food science has revolutionized the field by automating a broad range of applications such as food authenticity and food fraud detection. However, in order to create effective and general models able to address the complexity of real life problems, a vast amount of training samples are required.

Generally speaking, in machine learning and chemometrics, the bigger and more varied the calibration dataset is, the more accurate can be its classification power [1]. This number can be varied from a few dozens to many hundreds depending on the required accuracy [2]. This demand is increased even more due to the choice of specific pattern recognition methods that required a balanced dataset across classes to work effectively [3] or big number of samples in order to converge to a solution or optimize all their internal parameters [4]. Latest advances in deep learning [5,6] have massively overtook previous state-of-art methods by training neural networks with increasingly larger datasets. While simple classification and regression problems under controlled conditions can still be solved with limited training data, they tend to generate overfitted models to the particular training set and therefore not be generalised well to different experimental setups or real-world conditions, where their performance drastically falls. As a consequence, there is an increasing demand for larger and varied admixtures/samples datasets. Nevertheless, acquiring a diverse amount of samples is a time consuming and costly process, in which collecting samples representative of the real-world variation is not always possible.

In the field of food adulteration detection, this challenge is even more obvious. Sourcing pure and authentic commodities as well as adulterants in order to construct the models can be a very challenging task [7] and the official, and informal sources of true authentic samples (e.g. a rare spice or an exotic oil) are limited. This often results in studies with limited variability and overfitted models. In addition, to detect adulterants, current practice is to produce an appropriate number of in-house admixtures by mixing several commodity samples with one or more adulterants in different concentration grades. This allows for a robust classification/quantification model, but the number of combinations to be covered may become intractable. The preparation and the analysis of these samples require a lot of time, labour and other resources. Laboratory efforts have been made to simulate and approach mildly refined [8] or degraded samples [7]. Nonetheless, these approaches barely mitigate the problem [7,9] and still demand time consuming and expensive processes [8].

Assuming enough available samples, their characterisation through spectroscopic or chromato-

graphic methods is not absent of limitations towards the generality of the models. Thus, most chemometric methods described in the literature as well as commercial calibration models are based on data acquired by a single analytical instrument. This translates into models dependant on the spectroscopic instruments used for the data acquisition. The performance of those models with samples analysed by instrument from a different manufacturer is largely unknown and by experience unsuccessful [10]. Making the model "instrument agnostic" will require a multiplicity of instruments under various instrumental conditions so this variability can be incorporated through training into the model. This, however, can be impractical, make the cost unbearable and increase the time scale of a project. One practical solution could be to build and maintain spectral libraries on a higher performance laboratory instrument and transfer to other spectrometers using standardization protocols [10–12]. However, transferring calibration models from instrument to instrument is demanding because it requires the absorbancies/intensities of each feature in a set of selected samples obtained on the master instrument to be regressed against the corresponding absorbancies/intensities on the slave instrument.

In this study, a novel data augmentation solution is presented in order to efficiently mitigate previously mentioned problems and to obtain generalised classification models not only for a single instrument/lab validation but also for inter-lab and multi-instrument validation. Vegetable oils and spectroscopic data acquisition have been used here as a case study to demonstrate the influence of this innovative approach in chemometrics.

1.1. State of the art

The term data augmentation refers to methods for building more accurate, tolerant and flexible classification models via the introduction of unobserved data or latent variables [13]. Data augmentation has been widely applied in other machine learning application fields such as video processing [14], biometrics [15] or text analysis [16] to name a few, but very scarcely in chemometrics and food/analytical science. In order to avoid confusions, this term has to be differentiated from the data matrix augmentation where other experimentally measured data matrices under different conditions are appended (row-wise, column-wise or both) to introduce a new data structure [17].

Data augmentation methods have been applied to multivariate calibration of spectroscopic data in order to add sample variability for a single lab validation so far. These methods were mainly based on various types of 'noise' addition, otherwise referred as noise adaptation, to the original data set before calibration [18,19], aiming to represent some of the possible variations in real spectral data. Conlin et

al.[18] added Gaussian noise, with different levels of standard deviation, to the calibration set of NIR spectroscopy data spectra for a partial least squares (PLS) predictor. As a result, some improvement on the accuracy was achieved for the calibration models generated from the noise augmented data sets against those obtained solely from the original data set. Further studies have been conducted where noise augmentation methods were combined with ensemble methods[20]. Ensemble methods generate multiple chemometric models calibrated on independent noise-augmented training data and combine these to get an aggregated decision[20]. Bagging (bootstrap aggregating)[21,22] and boosting[23] are the most well known ensemble methods. Specifically, NIR spectroscopic data of vinegar samples were modeled with ensemble PLS and noise augmentation (additive noise, multiplicative noise, intensity-dependent noise, local-shift, instrumental noise or combination of them) for simulating the detection of possible fraudulent dilutions[24]. It was found that ensemble PLS models trained on augmented data led to calibration models presenting slightly better accuracy and robustness on the test set against possible perturbed new samples than ensemble PLS models on the original data only. However, all these previous attempts do not fully exploit the potential of data augmentation, by reducing it to the simple addition of noise to the raw spectra.

Other studies targeted specific formats of variation to be simulated. In these studies, the simulated noise is based on prior knowledge of the data variation. In a study[19], NIR spectra were noise augmented for improving the prediction of active pharmaceutical ingredient (API) in tablets using PLS regression. Noise augmented spectra were generated by adding the mean-centred spectra of the physical variations and unknown chemical variations (e.g. water content), which were calculated using orthogonal projection of pure component spectra, to the original calibration spectra. Segtnan et al.[25] added simulated noise on the spectra to handle temperature shifts regarding calibration dataset. The augmented NIR spectra were created through simulations based on experimental spectral data obtained at different temperatures (prior knowledge)[25]. Wavelength calibration errors and shifts, baseline offsets, path length changes, high levels of stray light, heteroscedastic noise, background contributions, multiplicative variations and variation in water content have been simulated in order to reduce overfitting to the NIR calibration set, as well as to optimise their parameters, in different learning methodologies as neural networks[26], PLS models[27] and principal component analysis[28]. These variations have been also introduced in the testing phase by generating artificially-derived test sets to better evaluate the tolerance of a model in different conditions and instrumental settings[29]. All aforementioned methods[19,25–29] have exhibited interesting results for multivariate regression and classification problems

on spectroscopic data due to the benefit of data augmentation methods. However, the augmentation techniques were closely linked to the application problem and require precise knowledge of the data in order to introduce the specific variation, which may impact their extension to other food science problem or spectroscopic methodologies.

To prevent the degradation of the performance of the calibration models due to unforeseen variations in spectra, different calibration maintenance and transfer methods have been proposed using augmented spectra. Haaland and Melgaard proposed the prediction-augmented classical least-squares (PACLS) where unmodeled spectral variations can be incorporated to classical least-squares (CLS) or PLS calibration models during the validation step[30,31]. An augmentation experiment was conducted using the Kennard-Stone subset selection algorithm where measured spectra augmented to simulated datasets to incorporate more spectral variability[32]. In another study, Haaland introduced a synthetic procedure in which quantitative spectral models with constant-temperature samples are augmented with a CLS estimate of the spectral effect of unmodeled temperature variations obtained from variable-temperature aqueous samples[33]. Moreover, to maintain the predictive abilities of a calibration model, Systematic Prediction Error Correction (SPEC) was developed for the cases where the spectroscopic instrument or measurement conditions are changed where only a few standardisation spectra are required for its application[34]. Tikhonov regularization (TR) has been used for updating a calibration model in order to predict samples acquired in different instruments[35,36], different temperatures[35] and different geographical regions[37]. Kramer and Small proposed a blank augmentation protocol as a modeling technique for the analysis of physiological levels of glucose[38]. Calibration transfer and maintenance to all the aforementioned studies were performed by augmenting the calibration model with only a few samples measured in the new secondary conditions.

Data augmentation has been used in process analytical technology (PAT)[39–42] which is a manufacturing concept, originated from pharmaceutical industry, where sources of sample variability are accounted and the production process is fitted to include this variability to improve the final product quality. One example where this concept is well shown includes a study[43], in which calibration laboratory samples (i.e. NIR spectra) are produced with the same physical variability as the production samples in pharmaceutical analysis for the determination of the API concentration by using a similar granulation treatment to the one used in industry.

More powerful data augmentation can be achieved by not only manipulating each sample in isolation but also exploiting the relationships among samples. A clear example is the generation of artificial

samples. In the literature, synthetic NIR calibration spectra were generated by convoluting measured background spectra with pure-component absorbance spectra for the determination of physiological levels of glucose in measured testing samples[32]. An important prerequisite of this strategy is a stable instrument or experimental setup. However, only minor preliminary attempts have been done in food authentication studies for the compositional evaluation of multi-varietal food blends. To cover the absence of a representative dataset simulating binary blends, artificial oil blends were generated combining the individual chemical indices of two different olive oil cultivars in various mixing ratios[44]. This need was arguably due to the chosen neural networks-based methodology that needs a larger dataset. Nevertheless, no validation of its influence in the final result is discussed. Semmar and Artaud[45] simulated a complete set of possible blends combining three olive oil varieties by using a simplex mixture design for the preparation of a broad data library in order to predict proportions of different co-occurring oil varieties in different blends using the chromatographic profile of the blend. The binary and ternary mixtures in varied proportions produced were characterized by the average fatty acid (FA) profiles calculated by combining the individual profiles.

In this paper we propose a novel data augmentation framework that generalises previous preliminary attempts in the literature and allows the introduction of not only noise augmentation techniques but other augmentation techniques such as artificial data blends or simulated acquisition instruments. By generalising and extending the concept of the data augmentation in the field of chemometrics, we aim to better handle the variation produced by different manufacturer instruments, inclusion of the human factor in data preparation and/or unbalanced training datasets.

2. Theory

2.1. The proposed data augmentation generator

Differences in the spectral acquisition of a given sample can be caused by sample preparation effects, instrumental drifts or other changes[24] that can affect considerably the classification performance and the stability of the chemometric models. To accurately predict the sample properties from spectra measured on different spectroscopic instruments than the one used to build the calibration model, an extra variance is also required for covering the various conditions of these secondary instruments.

Bearing in mind these different types of variability required, we designed and implemented a novel general framework for the application of the data augmentation techniques to spectra (see Figure

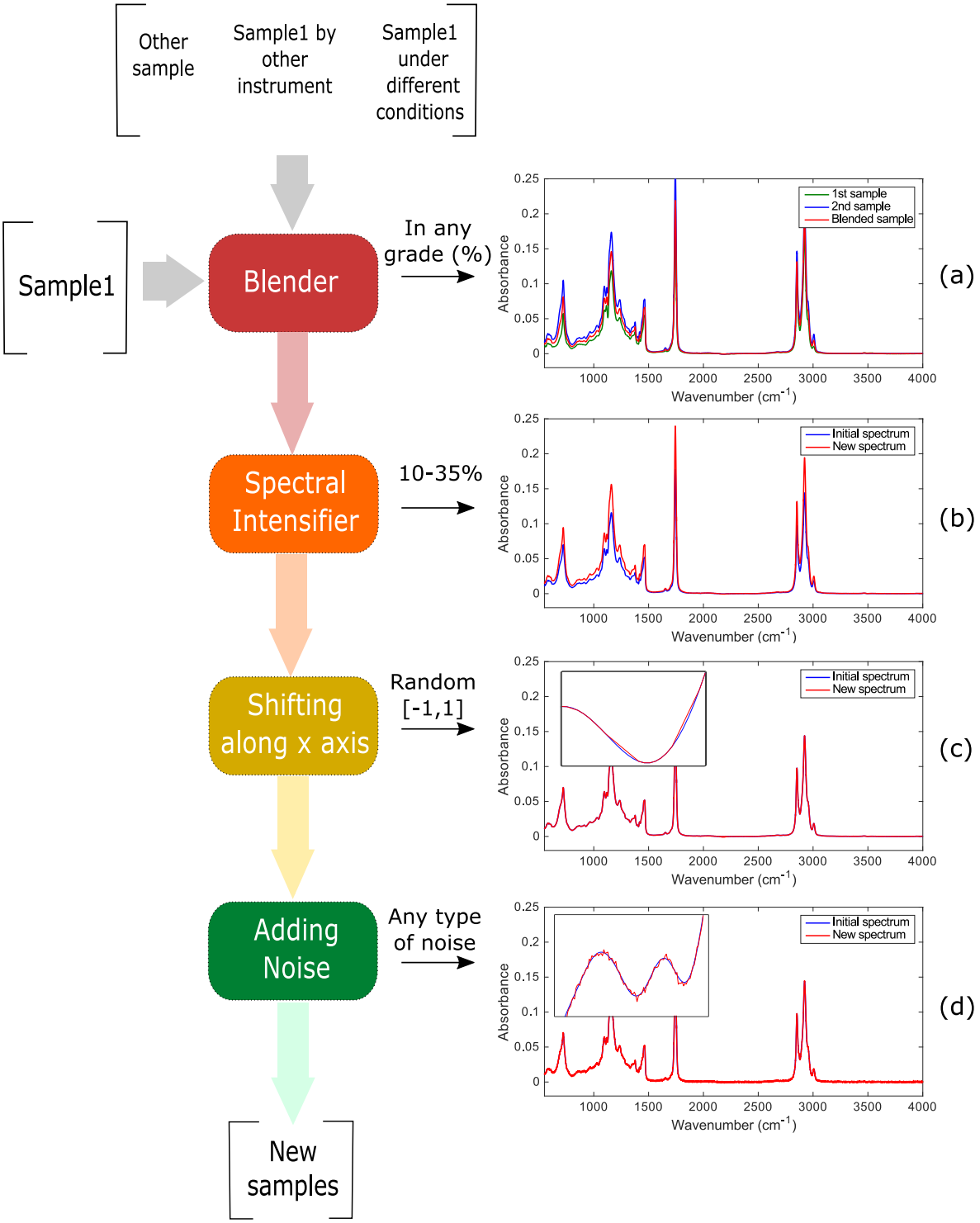

Figure 1. Scheme of the data augmentation framework.

1). This is a carefully designed pipeline of four data independent blocks which can be finely tuned depending on the desired variance for enhancing model's robustness: a) blending spectra, b) changing the intensity, c) shifting along x axis, and d) adding noise. Each of the four blocks can be enabled either alone or in combination with the others. The blocks also have input parameters that allow to be applied in higher or lower degree depending on the expected variability in testing. When the spectrum of a sample is acquired (here FT-IR spectroscopy), it is passed to our data augmentation framework where one or more samples (augmented samples) are generated from this particular one. The resulting original and augmented samples will then be passed to the chosen classification pipeline, where are preprocessed and used to calibrate the chemometric model.

Blender: The first component of our data augmentation scheme is the blender. This block aims to combine samples showing some variation in order to create artificial admixtures/samples as a weighted sum of the input samples. One input can be a sample, while the other inputs can be a different sample or the same sample acquired by different instrument (see Figure 1a) or under different conditions depending on the food classification problem examined. The new sample(s) are created by applying the weighted average of the input samples. Specifically, for generating an artificial admixture of m different input samples with n wavenumbers, their absorbances for a specific k wavenumber, $A_{1,k}$, $A_{2,k}$, $\ldots$, $A_{m,k}$, have to be multiplied by their concentration grades, per_1, per_2, $\ldots$, per_m:

$$Admixture_{new} = \{A_{new,k}\}_{(k=1..n)} = \frac{\sum_{i=1}^{m} per_i * A_{i,k}}{\sum_{i=1}^{m} per_i} \tag{1}$$

The intervals of the concentration grades for each input sample can be selected as equally distributed weights, i.e. $per_i = \frac{1}{m}, \forall i \in [1, m]$, as manually defined by the user, or as automatically defined to generate a range of admixtures between two reference i and $\tilde{i}$ with a given defined resolution r, so $per_i = \{x \in [0, 1] \mid x' = x + r\}$ and $per_{\tilde{i}} = 1 - per_i$. Using the Eq. (1), numerous artificial admixtures can be created for the study of an adulteration/contamination problem, a quantification problem and even for the creation of an instrument agnostic predictor (by combining the spectra of two or more instruments capturing the same sample).

Spectral intensifier: The second component allows modifying the intensity of a spectrum. Although many scattering effects in testing samples are corrected by the pre-treatment step, some of them are inevitable with result the misclassification of the samples. One of these cases is the baseline variations produced by the light scattering from spectra obtained by different spectroscopic instruments. To control this effect, new samples based on the real spectra can be generated (see Figure 1b)

by changing the absorbance of each k variable, $A_{original,k}$, of a spectrum.

$$Sample_{new} = \{A_{new,k}\}_{(k=1..n)} = (M * A_{original,k}) + C \tag{2}$$

where M is the amplification factor and C the baseline factor. Depending on the values of these two factors, the operation performed by this block is different. For values $M \neq 0$ and $C = 0$, an amplified version of the original spectrum is produced (multiplicative baseline offset), whereas a change of the baseline of the original spectrum occurred for $M = 1$ and $C \neq 0$ (baseline offset).

Shifting along x axis: As a third method, random shifting is applied along data points of the spectra for mimicking the instrumental variations. Randomly selected variables of a sample are shifted horizontally. The shift can be positive or negative which means a forward or backward shifting of the value of a variable respectively. New shifted augmented samples are subtly different from the original spectra (see Figure 1c).

$$A_{new,\dot{k}} = A_{original,k} \quad \text{being} \quad \dot{k} = k + round(L(b)) \tag{3}$$

where L a Laplacian distribution with a scale parameter b and location parameter $\mu = 0$. Laplacian distribution is chosen over other distributions such as Gaussian or uniform to ensure a very limited amount of shifting is generated and so the resulting spectrum is not unrealistic, since the shifting in the x-axis is not a common phenomenon.

Adding noise: Finally, the variability of a class in a training dataset can be increased by including these slightly noisy spectra based on the original spectra. Now, the new absorbance of each k variable, $A_{new,k}$, is the sum of the original absorbance $A_{original,k}$ and the noise w.

$$Sample_{new} = \{A_{new,k}\}_{(k=1..n)} = A_{original,k} + w \tag{4}$$

For generality, this added noise, w can be white Gaussian noise in specific signal-to-noise ratio per spectrum, in dB, which specifies the intensity of noise in this block (see Figure 1d). The addition of Gaussian noise to the original data has been proved to lead to calibration models with improved accuracy and enhanced robustness[18].

Generally speaking, all the described techniques of the proposed data augmentation scheme can derive a satisfactory number of samples for a class with very small number of original and representative samples for producing a balanced classification model. The different blocks have been designed to be as

general as possible independently of the data and/or application. Ranges for the values of parameters for each of these augmentation methods are fully configurable to adapt to different problems and applications. For our case of study, the identification of vegetable oil species in oil admixtures, chosen range values are indicated in Figure 1.

Both the dataset and the code of the data augmentation generator and chemometric data preprocessing will be available in the web for public use. Code was implemented using Matlab routines (The MathWorks Inc., USA).

3. Experimental

To evaluate the data augmentation mechanism, a previous study setup[46] was used as a base for this evaluation. The rapid identification of vegetable oil species[47] is used in this paper as case of study to prove the potential of data augmentation. In this application, 6 different classes, comprised of 3 pure oil types and their corresponding binary admixtures, should be distinguished.

First, the potential of our data augmentation generator will be demonstrated by augmenting a dataset and showing the improvement obtained regarding the same system without the artificial samples. Second, the blender is also validated independently. Then a batch of experiments are performed in a more complex setup, where multiple FT-IR instruments are used during the acquisition, with no overlap between the instruments used in training and testing. Correct classification rate[48] is used in all experiments as main evaluation metric.

3.1. Intra-laboratory experiment

3.1.1 Samples

Twenty refined vegetable oils were sourced from authentic palm oil and its derivatives (e.g. whole palm oil, palm stearin and palm olein) (PO), palm kernel (PKO), sunflower oil and rapeseed oil samples (see Table A.I in Appendix A). Binary admixtures were prepared in-house in different concentration grades from 16% to 84%. In total, 142 binary in-house admixture samples were included (n=162 samples including the twenty pure vegetable oils). Given the similarity of some of these oil samples and following the design of our previous study[46], we will consider rapeseed oil and sunflower oil equivalent and belonging to the same class. Thus, the classes to be identified are three pure (class 1 to 3) and 3 mixed classes (4 to 6): class 1 = PO; class 2 = RS (Rapeseed oil, Sunflower oil, Rapeseed

oil+Sunflower oil); class 3 = PKO; class 4 = RSPKO (RS+PKO); class 5 = RSPO (RS+PO); class 6 = PPKO (PO+PKO).

3.1.2 FT-IR spectral acquisition

The acquisition of all FT-IR spectra was performed using a Nicolet iS5 FT-IR spectrometer (Thermo Fisher Scientific, Dublin, Ireland) equipped with a DTGS KBr detector and a KBr beam splitter. Spectra were acquired from 4000 to 550 cm^{-1} co-adding 32 interferograms at 4 cm^{-1} resolution and a zero filling factor of 2 with a diamond attenuated total reflectance (iD5 ATR) accessory. Zero filling factor determines the number of levels of zero filling used when the data are Fourier transformed and therefore improves the line shape of a spectrum. At each spectrum point, absorbance values were recorded. Three replicates were acquired with initial 7157 data points and used in our experiments.

3.1.3 Data pre-treatment

The resulting FT-IR data underwent some pre-processing techniques to decrease or eliminate any random or systematic variation in the spectra[49]. Specifically, prior to the development of the multivariate models, Standard Normal Variate (SNV)[50], first order derivative[51], Savitzky-Golay filter[52] [polynomial order=2,frame size=9] and Pareto scaling[53] were applied for removing the scatter, correcting the baseline, smoothing the data points and scaling the data for preventing the dominance of high absorbances respectively. As a last step of the pre-processing procedure, the irrelevant spectra area was cut out. In total, 3781 variables between 654.23 and 1875.43 cm^{-1} and between 2520.02 and 3120.74 cm^{-1} were selected. All the aforementioned pre-processing techniques were selected empirically for the specific case study of the identification of vegetable oils.

3.1.4 Classification model

Soft modelling of class analogy (SIMCA)[54] as the modelling method and partial least squares discriminant analysis (PLS-DA)[55] as a discriminant method were used for identifying vegetable oil admixtures for this experiment.

3.1.5 Effect of data augmented samples for improving the performance of a chemometric model

This experiment was conducted in order to prove how useful are the data augmented spectra. In this

experiment, the classification performance is compared against the same system when data augmentation is added to the training set. Cross validation (venetian blinds) has been used as the evaluation method of the classification models, in which 1/7 of the samples is predicted with the remaining 6/7 of the samples and this procedure is repeated 7 times. The mean classification rate and the standard deviation over these iterations are the main evaluation metrics of this comparative analysis. For the augmented framework, each block is applied separately to the mean spectrum of each class. As a result, 216 augmented samples in each iteration, 36 per each class, were generated. SIMCA and PLS-DA parameter values were optimised to provide the best accuracy in the non-augmented case -see Table A.II at Appendix A for the exact numerical values- and kept fixed for the augmented set. The data augmentation parameter values were determined experimentally: Spectral intensifier: M=1.01-1.15 with a step of 0.01, C=0, shifting along axis: Laplacian distribution with b=0.6, and noise: Gaussian noise 38dB.

Furthermore, an additional experiment was performed to validate the spectral blender in isolation. With this aim, all admixture models in RSPKO, RSPO and PPKO classes were replaced by synthetic admixture samples generated by the blender from the pure oil samples for the calibration of the model. Synthetic samples were generated in the exact same concentration grades as the real in-house samples, producing 106 artificial admixtures. The classification ability of the model trained with artificial admixtures was compared against the performance of an equivalent model trained with the real lab admixtures. Both SIMCA and PLS-DA were used as classifiers in this testing.

3.2. Inter-laboratory experiment

This experiment involves the use of several instruments used to acquire the oil spectra. Therefore, this relevant experiment aims to simulate a more realistic environment where the model is not so closely related to the data acquisition. For this purpose, a trial with seventeen instruments including our laboratory instrument has been performed. These instruments belong to representatives of research centres, public services and private food testing labs (see Table A.III in Appendix A).

A total of nine (9) samples including pure oils and oil admixtures were prepared in our lab and sent to the participants having the instruments to collect the spectra. The oils used for the preparation of the inter-lab samples were from different geographical origin (Thailand) and year of production from the ones included in the calibration set. The pure oil and oil admixture samples were: Sample 1: 100% Palm oil (PO); Sample 2: 100% Rapeseed oil (RS); Sample 3: 100% Palm kernel oil (PKO); Sample

4: 50% Rapeseed + 50% Palm oil (RSPO); Sample 5: 70% Rapeseed + 30% Palm stearin (RSPO); Sample 6: 40% Palm kernel oil + 60% palm oil (PPKO); Sample 7: 50% Rapeseed oil + 50% Palm kernel oil (RSPKO); Sample 8: 40% Rapeseed oil + 60% Sunflower oil (RS); Sample 9: 70% Palm olein + 30% Rapeseed oil (RSPO). These samples will be used to validate and test a model calibrated with the samples described in Section 3.1.

3.2.1 FT-IR spectral acquisition

The acquisition parameters have been harmonised so that they are compatible with every FT-IR instrument. Linear interpolation was applied to spectra (n=126) from different instruments in order to get the desirable number of variables.

3.2.2 Data pre-treatment

The same pre-treatment techniques used in intra-laboratory experiment were employed in order to pretreat the inter-lab spectra (see Section 3.1.3).

3.2.3 Classification model

PLS-DA was used as classifier in these experiments since it is the most used discriminant supervised chemometric technique (commercial software and in-house routines) and its superior performance is demonstrated in Table I. For comparison purposes, the number of PLS-DA latent variables were optimised for the non-augmented set and selected to be the same for all the scenarios (PLS-DA: Lv=2).

3.2.4 Validation of Data augmentation in inter-lab trial

Two experiments are related to the inter-lab validation and how results can be improved using the data augmentation framework. The performance of the models without any data augmented spectra were presented and compared with the data augmented models.

Both scenarios handled the problem using three different datasets, calibration, validation and testing dataset. In the first scenario, the main dataset plus the spectra of the nine vegetable oils of the inter-lab trial acquired with the same spectrometer (our lab spectra) were used for the training of the model. Five out of 16 remaining instruments were randomly selected as a validation dataset for tuning the different parameters of our data augmentation system. The final model resulted was tested

with the remaining instruments (eleven instruments). Results are compared against the same pipeline without the data augmentation methods. This experiment aims to show how data augmentation helps to make results more general and robust against undesired variability such as the one introduced by the capturing instrument. The input values for the blocks of the data augmentation solution applied to the mean spectrum of each class and chosen experimentally after validation were: Spectral intensifier: M=1.01-1.35 with a step of 0.01, C=0, shifting along x-axis: Laplacian distribution with b=0.6, noise=25dB Gaussian noise. As a result of these input values, 76 augmented samples were produced and added for each class.

Finally, in the second inter-lab experiment, the two instruments producing the most extreme spectra (9^{th} and 13^{th} instruments in Table A.III in Appendix A) were selected for the improvement of the classification by applying the weighted average of the inter-lab samples of these two different instruments. This sub-experiment aims to demonstrate how the blender component can be used to simulate an infinite amount of variability due to instrumentation that can be thus incorporated to our model to make it even more robust against it. Specifically, artificial samples were generated by combining the same samples produced by these two instruments in varied concentration grades. Thus, the original samples from our spectrometer and the new artificial samples from the virtual instruments were added to the main dataset for the calibration step. Similarly to the previous sub-experiment, four instruments were applied for validation and parameter tuning and the remaining instruments (ten instruments instruments) for testing. The input values for the blocks of the data augmentation solution performed on the mean spectrum of each class and selected empirically after validation were: Spectral intensifier: M=1.01-1.33 with a step of 0.01, C=0, shifting along x-axis: Laplacian distribution with b=0.6, noise: 35dB Gaussian noise. Specifically, 64 augmented samples were created for each class. Regarding the blending of the two instruments, 35 new artificial samples produced by the blender for each inter-lab sample (instrument weights *per* from 16% to 84% with step 2%).

4. Results and discussion

As a result of the different scenarios proposed in Section 3, the application of the proposed data augmentation scheme has been assessed on three main different experiments and its outcomes have been compared to those obtained without any data augmentation technique. These experiments assess how data augmentation scheme can enhance classification results.

4.1. Intra-lab validation

4.1.1 Experiment 1: Effect of data augmented samples for improving the performance of a chemometric model

In this experiment, we aim to demonstrate that the introduction of data augmented spectra improves the performance of the chosen pipeline. Specifically, the augmented training set increases the performance of the learned model up to 19% the mean classification rate on real samples, depending on the classification technique, and reduces the standard deviation compared with a model trained on actual samples only (see Table I). A significantly bigger improvement is achieved for SIMCA than with PLS-DA, but this is mainly due to the lower performance baseline which has more space for improvement, rather than due to any limitation of our augmented framework to be combined with PLS-DA, as we will show in the experiments 2 and 3. In any case, the improvement in PLS-DA is also clear since, not only the average accuracy improves, but also the standard deviation reduces. Similar results, with only negligible differences, were obtained when the average of the three replicates is used instead of using all the replicates (Results not shown).

Table I. Mean classification rate (%) and standard deviation in validation using non-augmented and augmented calibration models. Cross validation was applied (venetian blinds).

Data augmented samples in the calibration of a model		
Classification technique	Only actual lab samples in training	Actual lab + artificial samples in training
SIMCA	64 ± 2.4	77 ± 4
PLS-DA	98 ± 1.4	99 ± 1

With respect to the validation of the blender, the system trained with artificial admixture gave almost identical results than the same system trained with real admixtures, being the former only 5.66% smaller in average than the later. Given that the standard deviation reported in Table I, this difference can be consider small taken into account that no real admixture was used at all in the calibration. The outputs of the blender can therefore be considered realistic.

4.2. Inter-lab validation

4.2.1 Experiment 2: Data augmentation without virtual instrument simulation

Table II shows the improvements of the data augmentation in this scenario. First of all, it can be noticed the performance drops from almost 100% to ∼60% in spite of using the same chemometric

pipeline than in the intra-lab experiments. This major decrease corroborates the more complex and realistic problem when validating using multiple different instruments. Under these conditions, the different data augmentation blocks increasingly improve the performance of the learned model. The use of data augmentation methods produces a more robust and generalised classification model. The very good classification behaviour produced by the validation step is retaining in the testing step.

Table II. Classification rate(%) for each of the following cases of data augmentation generator for the testing and validation step by using PLS-DA (Lv=2) using one participant for the training

Data augmentation technique	**Classification rate (%)**
Validation step	
Without data augmentation	58
Spectral intensifier (M=1.01-1.35)	73
Spectral intensifier (M=1.01-1.35) + Shifting along x-axis (b=0.6)	78
Spectral intensifier (M=1.01-1.35) + Shifting along x-axis (b=0.6)+ Gaussian noise (25dB)	82
Testing step	
Without data augmentation	61
Spectral intensifier (M=1.01-1.35) + Shifting along x-axis (b=0.6)+ Gaussian noise (25dB)	74

Figure 2c shows the projection of inter-lab testing samples on the PCA space of the training dataset. Figure 2a and 2b show the PCA space of the calibration data set before and after the application of the data augmentation methods. It can be observed how the data augmented samples increase the variation of each class separately and better cover the amount of variability in the testing spectroscopic data, caused by the use of different FT-IR configurations (different types of ATR sample module, varied detectors from manufacturer to manufacturer, etc.) and different users (technical vs non-technical users in different organisations).

4.2.2 Experiment 3: Data augmentation with virtual instrument simulation

In the second sub-experiment, in addition to the previous components of the data augmentation, the blender component has been used for blending the spectral data produced by two instruments in the inter-lab trial. Results in Table III demonstrate how data augmented samples can improve classification rate by a 38.61%, validating our approach and the potential of the blender to simulate variability between instruments that is successfully incorporated into the model.

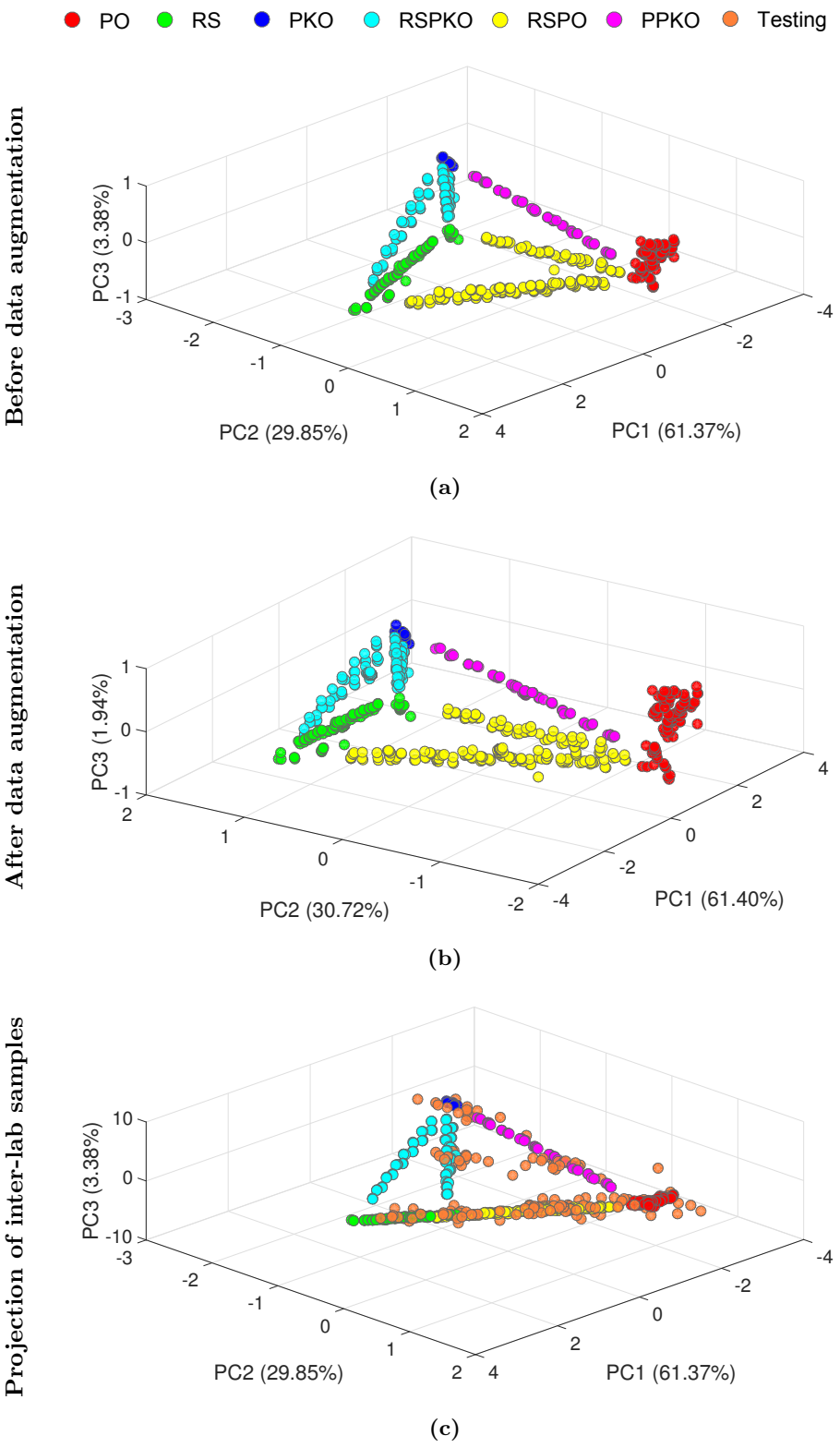

Figure 2. PCA exploratory analysis of training data: (a) space before data augmentation for 1st scenario of inter-lab trial validation; (b) space after data augmentation for 1st scenario of inter-lab trial validation; ; (c) Projection of all inter-lab samples (orange) on the PCA space of original training data.

Table III. Classification rate(%) for each of the following cases of data augmentation generator for the testing and validation step by using PLS-DA (Lv=2) using three participants for the training

Data augmentation technique	Classification rate (%)
Validation step	
Without data augmentation	67
Spectral intensifier (M=1.01-1.33)	75
Spectral intensifier (M=1.01-1.33) + Shifting along x-axis (b=0.6)	81
Spectral intensifier (M=1.01-1.33) + Shifting along x-axis (b=0.6)+ Gaussian noise (35dB)	83
Spectral intensifier (M=1.01-1.33) + Shifting along x-axis (b=0.6)+ Gaussian noise (35dB) + mixing the samples of the two participants (16% to 84% with step 2%)	92
Testing step	
Without data augmentation	63
Spectral intensifier (M=1.01-1.33) + Shifting along x-axis (b=0.6)+ Gaussian noise (35dB) + mixing the samples of the two participants (16% to 84% with step 2%)	88

The PCA space of the calibration data set (Figure 3) indicates the result of this combination of data augmentation techniques and how the space between extreme real samples is covered by the blended artificial samples. The latent space produced retains its original structure but expanding on a third dimension where the variability of the virtual instruments is represented. This can justify the great performance of the model for both validation and testing steps.

5. Conclusions

In this paper, we have described a general data augmentation framework for chemometric analysis of spectral data aimed at those that develop methods for detection of food authenticity. Our solution generalised preliminary and basic approaches to data augmentation emerging in the field of chemometrics. This approach has been successfully validated on a case of study consisting on classifying vegetable oils using FT-IR spectroscopic data. The introduction of the data augmentation framework allows us to overcome the need to have big training data sets a priori. The augmented spectra were clearly beneficial in improving classification ability of a model (up to a maximum 19% improvement) in a qualitative study by introducing realistic variation through noise and displacements. Moreover, data augmented samples can enhance the robustness and generality of an instrument agnostic classification model by adding more variability not only among samples but also over the instruments (more than

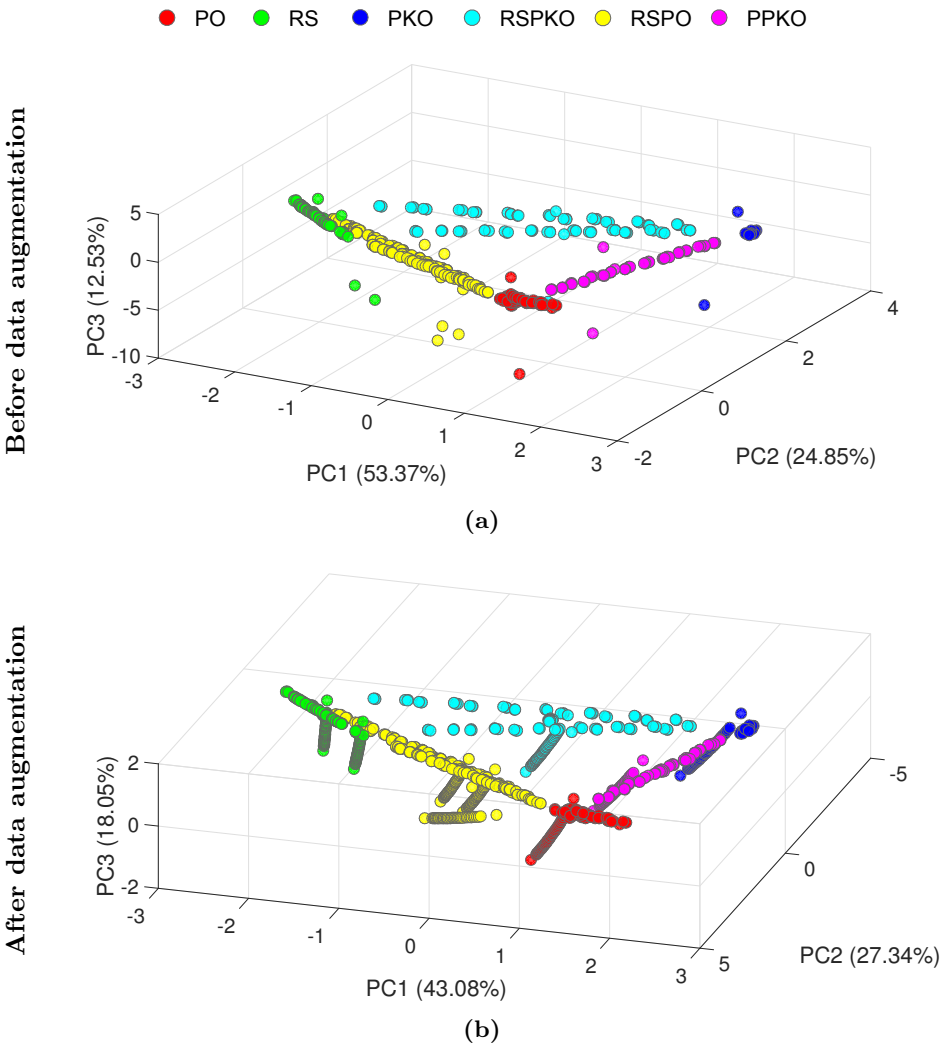

Figure 3. PCA exploratory analysis of training data : (a) space before data augmentation for 2nd scenario of inter-lab trial validation; (b) space after data augmentation for 2nd scenario of inter-lab trial validation.

38% improvement). The proposed data augmentation scheme does not affect or change the chosen pipeline by the designer and it is compatible with other methodologies to improve performance such as ensemble methods. It should be noted, however, that data augmentation is never a better substitute for real samples encompassing the true variability, but it may mitigate the need for as many samples and, importantly, could include sources of variability that would be difficult to achieve experimentally.

As future work, we plan to use the data augmentation pipeline to simulate and enhance cultivar variability, which was not possible in this paper due to the limitations of the current dataset regarding cultivars. In particular, we will validate the capacity of the blender to generate and simulate new

411 cultivars by combining oils from different origins. Moreover, we aim to demonstrate the generality of
412 our data augmentation framework by assessing it in a wider domain of food classification problems
413 and on other spectroscopic data like near infrared spectra as well as chromatographic data.

Acknowledgements

The authors would like to thank all the participants (research centres, public services and private food testing labs) that helped to perform the inter-laboratory experiment. This research was supported with funding from The Department Learning and Employment Northern Ireland (DELNI) (PhD studentship block grant) and the Department of Environment, Food and Rural Affairs (DEFRA) of the UK (Grant no. FAO 157).

A. Appendix

Table A.I. Details of pure samples

Pure Vegetable oil samples			
Species	**Identity of vegetable oil**	**Origin**	**Sample Provider**
Palm Oil (PO)	Whole palm oil 1	Not provided	Multinational consumer goods company
	Whole palm oil 2	New Britain island, Papua New Guinea	National sustainable palm oil refinery
	Whole palm oil 3	Not provided	Multinational consumer goods company
	Whole palm oil 4	Papua New Guinea	National sustainable palm oil refinery
	Whole palm oil 5	Indonesia / S. America	National sustainable palm oil refinery
	Palm stearin 1	Papua New Guinea	National sustainable palm oil refinery
	Palm stearin 2	Indonesia	Multinational provider of edible oils and fats
	Palm olein 1	Papua New Guinea / Malaysia	Multinational provider of edible oils and fats
Palm kernel oil (PKO)	PKO1	Not provided	Multinational consumer goods company
	PKO2	New Britain island, Papua New Guinea	National sustainable palm oil refinery
	PKO3	Not provided	Multinational consumer goods company
	PKO4	Papua New Guinea	National sustainable palm oil refinery
Sunflower oil	Sunflower oil 1	Not provided	Multinational consumer goods company
	Sunflower oil 2	Not provided	National oil supplier
	Sunflower oil 3	Not provided	Multinational consumer goods company
	Sunflower oil 4	EEC/France	Multinational provider of edible oils and fats
Rapeseed oil	Rapeseed oil 1	Not provided	Multinational consumer goods company
	Rapeseed oil 2	Not provided	National oil supplier
	Rapeseed oil 3	Not provided	Multinational consumer goods company
	Rapeseed oil 4	EEC/France	Multinational provider of edible oils and fats

Table A.II. Parameters empirically chosen for PLS-DA and SIMCA when using data augmentation or only real data

SIMCA							PLS-DA	
PCA	PO	RS	PKO	RSPKO	RSPO	PPKO	Lv	24
dimensions	20	20	3	20	20	20		

Table A.III. Instruments for the inter-lab validation of the classification model for the identification of vegetable oil species. N/a, not available

Id	**Participant**	**FT-IR Instrument**	**Detector**	**Year**
1	Teagasc, Food Research Centre	Bio-Rad Excalibur FTS 3100	DTGS	2001
2	PerkinElmer Ltd	PerkinElmer Spectrum 2	DTGS	2012
3	PerkinElmer Ltd	PerkinElmer Frontier	DTGS	2013
4	Brennan and Co.	Bruker Alpha	DTGS	2013
5	Public Analyst Scientific Services	PerkinElmer Spectrum 100	LiTaO3	2007
6	LGC Limited	PerkinElmer Spectrum One	DTGS	2001
7	Premier Analytical Services (Premierfoods)	Bio-Rad Excalibur FTS300MX	DTGS	2002
8	Institute of Food Research (IFR)	Nicolet MagnaIR 860	DTGS	1998
9	Institute of Food Research (IFR)	Bio-Rad FTS6000	DTGS	1996
10	Institute of Food Research (IFR)	Thermo Fisher Scientific Nicolet iN10MX/iZ10	DTGS	2011
11	Shimadzu (Mason Technology)	Shimadzu IRAffinity-1S	DLaTGS	n/a
12	Antech(IRE)	Thermo Fisher Scientific TruDefender FTX	DLaTGS	n/a
13	Agri-Food and Biosciences Institute (AFBI)	PerkinElmer Spectrum One	MIR TGS	n/a
14	Walloon Agricultural Research Centre (CRA-W)	Bruker Vertex 70	DLaTGS	2007
15	Walloon Agricultural Research Centre (CRA-W)	Bruker Vertex 70	DLaTGS	2012
16	Walloon Agricultural Research Centre (CRA-W)	Bruker Vertex 70	MCT	2012
17	Our lab (Institute for Global Food Security, Queen's University Belfast)	Thermo Fisher Scientific Nicolet iS5	DTGS	2012

References

1. Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Transactions on Pattern Analysis & Machine Intelligence 1991;(3):252–264.

2. Georgouli K, Del Rincon JM, Koidis A. Continuous statistical modelling for rapid detection of adulteration of extra virgin olive oil using mid infrared and Raman spectroscopic data. Food Chemistry 2017;217:735–742.

3. Berrueta LA, Alonso-Salces RM, Héberger K. Supervised pattern recognition in food analysis. Journal of Chromatography A 2007;1158(1):196–214.

4. Marini F. Artificial neural networks in foodstuff analyses: Trends and perspectives A review. Analytica Chimica Acta 2009;635(2):121–131.

5. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014;.

6. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.

7. Zhao J, Lin H, Chen Q, Huang X, Sun Z, Zhou F. Identification of eggs freshness using NIR and support vector data description. Journal of food Engineering 2010;98(4):408–414.

8. Perez-Camino MdC, Cert A, Romero-Segura A, Cert-Trujillo R, Moreda W. Alkyl esters of fatty acids a useful tool to detect soft deodorized olive oils. Journal of agricultural and food chemistry 2008;56(15):6740–6744.

9. Haugh R. A new method for determining the quality of an egg. US Egg Poultry 1937;39:27–49.

10. Grelet C, Pierna JF, Dardenne P, Baeten V, Dehareng F. Standardization of milk mid-infrared spectra from a European dairy network. Journal of dairy science 2015;98(4):2150–2160.

11. Bouveresse E, Massart D. Improvement of the piecewise direct standardisation procedure for the transfer of NIR spectra for multivariate calibration. Chemometrics and intelligent laboratory systems 1996;32(2):201–213.

12. Rodriguez JD, Westenberger BJ, Buhse LF, Kauffman JF. Standardization of Raman spectra for transfer of spectral libraries across different instruments. Analyst 2011;136(20):4232–4240.

13. Van Dyk DA, Meng XL. The art of data augmentation. Journal of Computational and Graphical Statistics 2001;10(1).

14. Stallkamp J, Ekenel HK, Stiefelhagen R. Video-based face recognition on real-world data. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on IEEE; 2007. p. 1–8.

15. Duval S, Tweedie R. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000;56(2):455–463.

16. Coates A, Carpenter B, Case C, Satheesh S, Suresh B, Wang T, et al. Text detection and character recognition in scene images with unsupervised feature learning. In: Document Analysis and Recognition (ICDAR), 2011 International Conference on IEEE; 2011. p. 440–445.

17. Tauler R, Maeder M, De Juan A, et al. Multiset data analysis: extended multivariate curve resolution. In: Brown SD, Tauler R, Walczak B, editors. Comprehensive Chemometrics, vol. 2 Oxford: Elsevier Science; 2009.p. 473–505.

18. Conlin A, Martin E, Morris A. Data augmentation: an alternative approach to the analysis of spectroscopic data. Chemometrics and intelligent laboratory systems 1998;44(1):161–173.

19. Pieters S, Saeys W, Van den Kerkhof T, Goodarzi M, Hellings M, De Beer T, et al. Robust calibrations on reduced sample sets for API content prediction in tablets: Definition of a cost-effective NIR model development strategy. Analytica chimica acta 2013;761:62–70.

20. Mevik BH, Segtnan VH, Næs T. Ensemble methods and partial least squares regression. Journal of chemometrics 2004;18(11):498–507.

21. Breiman L. Bagging predictors. Machine learning 1996;24(2):123–140.

22. Cheng J, Liu Q, Lu H, Chen YW. Ensemble learning for independent component analysis. Pattern Recognition 2006;39(1):81–88.

23. Tan C, Li M, Qin X. Study of the feasibility of distinguishing cigarettes of different brands using an Adaboost algorithm and near-infrared spectroscopy. Analytical and bioanalytical chemistry 2007;389(2):667–674.

24. Sáiz-Abajo M, Mevik BH, Segtnan V, Næs T. Ensemble methods and data augmentation by noise addition applied to the analysis of spectroscopic data. Analytica chimica acta 2005;533(2):147–159.

25. Segtnan VH, Mevik BH, Isaksson T, Næs T. Low-cost approaches to robust temperature compensation in near-infrared calibration and prediction situations. Applied spectroscopy 2005;59(6):816–825.

26. Gemperline PJ. Rugged spectroscopic calibration for process control. Chemometrics and intelligent laboratory systems 1997;39(1):29–40.

27. Despagne F, Massart DL, de Noord OE. Optimization of partial-least-squares calibration models by simulation of instrumental perturbations. Analytical Chemistry 1997;69(16):3391–3399.

28. Rutan SC, de Noord OE, Andréa RR. Characterization of the sources of variation affecting near-infrared spectroscopy using chemometric methods. Analytical chemistry 1998;70(15):3198–3201.

29. Wortel VA, Hansen W, Wiedemann S. Optimising multivariate calibration by robustness criteria. Journal of Near Infrared Spectroscopy 2001;9(2):141–151.

30. Haaland DM, Melgaard DK. New prediction-augmented classical least-squares (PACLS) methods: application to unmodeled interferents. Applied Spectroscopy 2000;54(9):1303–1312.

31. Haaland DM, Melgaard DK. New classical least-squares/partial least-squares hybrid algorithm for spectral analyses. Applied Spectroscopy 2001;55(1):1–8.

32. Sulub Y, Small GW. Spectral simulation methodology for calibration transfer of near-infrared spectra. Applied spectroscopy 2007;61(4):406–413.

33. Haaland DM. Synthetic multivariate models to accommodate unmodeled interfering spectral components during quantitative spectral analyses. Applied Spectroscopy 2000;54(2):246–254.

34. Chen ZP, Li LM, Yu RQ, Littlejohn D, Nordon A, Morris J, et al. Systematic prediction error correction: A novel strategy for maintaining the predictive abilities of multivariate calibration models. Analyst 2011;136(1):98–106.

35. Kunz MR, Kalivas JH, Andries E. Model updating for spectral calibration maintenance and transfer using 1-norm variants of Tikhonov regularization. Analytical chemistry 2010;82(9):3642–3649.

36. Kunz MR, Ottaway J, Kalivas JH, Andries E. Impact of standardization sample design on Tikhonov regularization variants for spectroscopic calibration maintenance and transfer. Journal of Chemometrics 2010;24(3-4):218–229.

37. Kunz MR, Ottaway J, Kalivas JH, Georgiou CA, Mousdis GA. Updating a synchronous fluorescence spectroscopic virgin olive oil adulteration calibration to a new geographical region. Journal of agricultural and food chemistry 2011;59(4):1051–1057.

38. Kramer KE, Small GW. Blank augmentation protocol for improving the robustness of multivariate calibrations. Applied spectroscopy 2007;61(5):497–506.

39. Pomerantsev AL, Rodionova OY. Process analytical technology: a critical view of the chemometricians. Journal of Chemometrics 2012;26(6):299–310.

40. van den Berg F, Lyndgaard CB, Sørensen KM, Engelsen SB. Process analytical technology in the food industry. Trends in food science & technology 2013;31(1):27–35.

41. Munir MT, Yu W, Young B, Wilson DI, Information I. The current status of process analytical technologies in the dairy industry. Trends in Food Science & Technology 2015;43(2):205–218.

42. Chen Z, Lovett D, Morris J. Process analytical technologies and real time process control a review of some spectroscopic issues and challenges. Journal of Process Control 2011;21(10):1467–1482.

43. Blanco M, Bautista M, Alcala M. Preparing calibration sets for use in pharmaceutical analysis by NIR spectroscopy. Journal of pharmaceutical sciences 2008;97(3):1236–1245.

44. Marini F, Magri AL, Bucci R, Magrì AD. Use of different artificial neural networks to resolve binary blends of monocultivar Italian olive oils. Analytica chimica acta 2007;599(2):232–240.

45. Semmar N, Artaud J. A new simplex-based approach predicting olive oil blend compositions from fatty acid data. Journal of Food Composition and Analysis 2015;43:149–159.

46. Osorio MT, Haughey SA, Elliott CT, Koidis A. Identification of vegetable oil botanical speciation in refined vegetable oil blends using an innovative combination of chromatographic and spectroscopic techniques. Food chemistry 2015;189:67–73.

47. Osorio MT, Haughey SA, Elliott CT, Koidis A. Evaluation of methodologies to determine vegetable oil species present in oil mixtures: Proposition of an approach to meet the EU legislation demands for correct vegetable oils labelling. Food Research International 2014;60:66–75.

48. Oliveri P, Downey G. Multivariate class modeling for the verification of food-authenticity claims. TrAC Trends in Analytical Chemistry 2012;35:74–86.

49. Nunes CA, Alvarenga VO, de Souza Sant'Ana A, Santos JS, Granato D. The use of statistical software in food science and technology: Advantages, limitations and misuses. Food Research International 2015;75:270 – 280. http://www.sciencedirect.com/science/article/pii/S0963996915300557.

50. Barnes RJ, Dhanoa MS, Lister SJ. Standard Normal Variate Transformation and Detrending of Near-Infrared Diffuse Reflectance Spectra. Appl Spectrosc 1989 May;43(5):772–777. http://as.osa.org/abstract.cfm?URI=as-43-5-772.

51. Osborne BG, Fearn T, Hindle PH, Hindle PT. Practical NIR Spectroscopy with Applications in Food and Beverage Analysis. Longman food technology, Longman Scientific & Technical; 1993. https://books.google.co.uk/books?id=0XV1QgAACAAJ.

52. Savitzky A, Golay MJ. Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry 1964;36(8):1627–1639.

53. Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC genomics 2006;7(1):1.

54. Wold S. Pattern recognition by means of disjoint principal components models. Pattern recognition 1976;8(3):127–139.

55. Barker M, Rayens W. Partial least squares for discrimination. Journal of chemometrics 2003;17(3):166–173.